(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
6 September 2002 (06.09.2002)

PCT

(10) International Publication Number
WO 02/068579 A2

(51) International Patent Classification[7]: C12G

(21) International Application Number: PCT/US02/00284

(22) International Filing Date: 10 January 2002 (10.01.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/756,696 10 January 2001 (10.01.2001) US

(71) Applicant: **PE CORPORATION (NY)** [US/US]; 761 Main Avenue, Norwalk, CT 06859 (US).

(72) Inventors: **VENTER, Craig, J.**; Celera Genomics, 45 West Gude Drive C2-4#21, Rockville, MD 20850 (US). **ADAMS, Mark**; Celera Genomics, 45 West Gude Drive C2-4#21, Rockville, MD 20850 (US). **LI, Peter, W., D.**; Celera Genomics, 45 West Gude Drive C2-4#21, Rockville, MD 20850 (US). **MYERS, Eugene, W.**; Celera Genomics, 45 West Gude Drive C2-4#21, Rockville, MD 20850 (US).

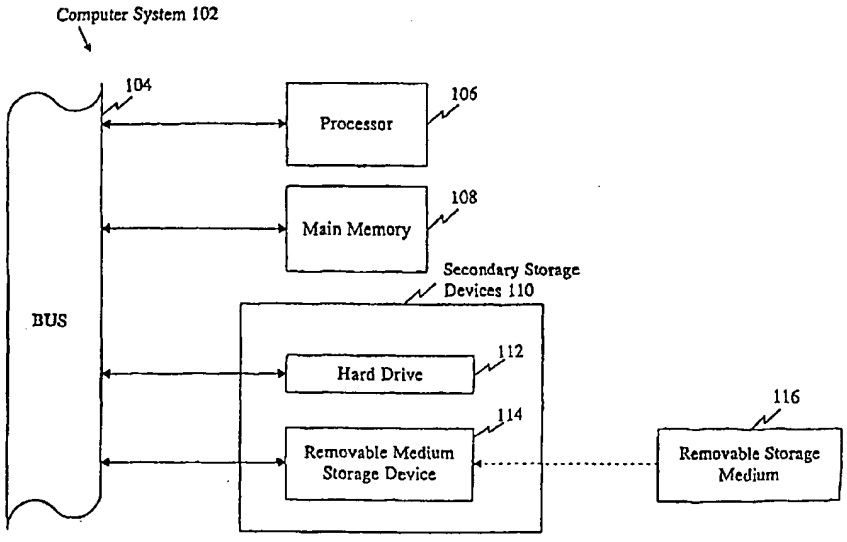
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: KITS, SUCH AS NUCLEIC ACID ARRAYS, COMPRISING A MAJORITY OF HUMAN EXONS OR TRANSCRIPTS, FOR DETECTING EXPRESSION AND OTHER USES THEREOF



(57) Abstract: The present invention is based on the sequencing and assembly of the human genome. The present invention provides the primary nucleotide sequence of the coding portion of the human genome in the form of a series of transcript sequences with accompanying exon information. This information can be used to generate nucleic acid detection reagents and kits such as nucleic acid arrays, and for other uses.

WO 02/068579 A2

KITS, SUCH AS NUCLEIC ACID ARRAYS, COMPRISING A MAJORITY OF HUMAN EXONS OR TRANSCRIPTS, FOR DETECTING EXPRESSION AND OTHER USES THEREOF

FIELD OF THE INVENTION

The present invention is in the field of genomic discovery systems. The present invention specifically provides the coding sequences of the human genome, including transcript sequences and corresponding exon information, in a form that is commercially useful, including detection kits and reagents such as nucleic acid arrays.

BACKGROUND OF THE INVENTION

The human genome is organized into discrete expression units called genes. Genes are further divided into exons (coding sequences) and introns (intervening, non-coding sequences). RNA transcripts are the primary output of the genome and are generated through a process referred to as gene expression or transcription. Gene expression involves the transcription of DNA into pre-mRNA, followed by RNA processing of pre-mRNA into mature mRNA transcripts, during which introns are removed and exons are spliced together to form complete transcript sequences. However, alternative splicing pathways allow introns to be removed and exons to be combined in different combinations, thereby allowing different mRNAs and proteins to be produced from the same gene. It has been found that nearly 40% of human genes are alternatively spliced (Brett *et al.*, 2000, *FEBS Lett.*, 474, 83). Different splice forms of genes may play distinctly different, and important, roles in different cells/tissues, developmental stages, or diseases and, therefore, the ability to detect different splice forms of the same gene is of paramount importance. Alternative splicing can also act as an on-off mechanism for mRNA activity by producing either functional or non-functional mRNAs from the same pre-mRNA.

A major goal in the development of therapeutics, diagnostic reagents, and pharmaceutical drugs is to understand and elucidate gene expression patterns and splicing patterns, particularly in different cells/tissues, developmental stages, and disease/pathological conditions. Determining when or under what conditions a particular gene or splice form is expressed, in which cells/tissues, and to what extent is important for understanding the function of the protein encoded by the gene and its role in disease. Gene expression and splicing patterns can be determined by reagents or kits, preferably nucleic acid arrays (also known as "DNA chips" or "biochips"), that utilize detection elements, such as nucleic acid probes, to detect the expression of gene fragments or the splicing together of exons to form mRNA transcripts. Such detection

elements may comprise, for example, fragments of, or complete, gene transcripts or exons, fragments corresponding to UTR regions of the transcript or detection elements that span the exon/exon boundaries of a transcript. The use of exons, or exon fragments, as detection elements has the distinct advantage of allowing the detection of different alternatively spliced transcript forms with the same detection element. This is possible so long as the transcript form contains the particular exon that is used as a detection element, regardless of how that exon is combined with other exons. On the other hand, the use of complete transcripts, or transcripts comprising more than one exon, generally allows the detection of only that particular splice form, or exon combination, and may not detect the expression of other important transcript splice forms. However, the use of transcripts as detection elements is advantageous in particular situations, such as when detection of only one particular transcript, with a high degree of specificity and minimal cross hybridization to other transcript forms, is desired.

The primary sequence of the exons/transcripts of the human genome would therefore be valuable for use in detection kits and reagents, such as nucleic acid arrays, for detecting gene expression patterns, including variable gene expression such as alternative splicing, and other uses. Human exons/transcripts can serve as detection elements, such as probes, in detection kits and reagents such as nucleic acid arrays. Not only will such kits and reagents serve as a basis for discovery and validation of commercially important genes, they provide commercially valuable tools for understanding the complex patterns of gene expression in relationship to different cells/tissues, developmental stages, and disease conditions. Consequently, human exons/transcripts, provided in a usable form, such as in the form of detection elements in a nucleic acid array, would be valuable for disease diagnosis and treatment, such as by improving the drug discovery and development process, or for diagnosing diseases based on aberrant gene expression patterns.

Furthermore, a substantial proportion of current gene discovery efforts is directed at mining EST databases. However, it has been estimated that EST databases may contain as little as 40% of the protein-coding portion of the human genome (Aparicio, *Nature Genetics*, June 2000, 25: 129-130). Consequently, the primary sequence of human transcripts and exons, identified through whole-genome sequencing, assembly, and annotation, represents the best source of identifying protein-coding sequences of the human genome that are not represented in EST databases. Therefore, the sequence of human exons/transcripts provided by the present invention is useful for identifying and validating commercially valuable human genes.

Gene expression analysis, using the transcript/exon sequences provided herein, is also useful for determining functions and relationships of genes with unknown functions. For

example, it has been shown in yeast that genes with similar functions have similar gene expression profiles (Eisen *et al.* (1998) *Proc. Natl. Acad. Sci. U.S.A.* 95, 14863-14868).

The present invention advances the art by providing the predicted transcript sequences (SEQ ID NOS:1-39010), for 39010 transcripts predicted from the assembled human genome, many of which did not have evidence for their existence in the prior art. Furthermore, the present invention provides information on each of the exons (Table 1) contained within the transcripts. The exon information contained in Table 1 includes the coordinates of each exon within its respective transcript, thereby allowing one to readily determine the precise boundaries of each of the exons using the transcript coordinates and the transcript sequences as a reference. These exon boundaries define the exon-exon junctions discussed herein. Also provided in Table 1 is evidence supporting the existence of each exon or transcript (e.g. EST hit, mouse hit, etc.).

Given the transcript sequences provided by the present invention and the exon coordinate information provided in Table 1, or fragments thereof, readily implementable compositions of matter, such as detection elements and detection reagent/kits, (e.g. in the form of probes in a nucleic acid array), can be made using methods well known in the art and discussed herein. Such kits and reagents can be used to track the expression and/or splicing of all of the transcripts/genes disclosed herein, the novel members herein provided, or rationally selected subsets thereof, defined by a user.

Nucleic Acid Arrays and Detection Kits and Reagents

Oligonucleotide probes have long been used to detect complementary nucleic acid sequences in a nucleic acid of interest (the "target" nucleic acid) in the form of detection kits and reagents. In some assay formats, the oligonucleotide probe is tethered, i.e., by covalent attachment, to a solid support, and arrays of oligonucleotide probes immobilized on solid supports have been used to detect specific nucleic acid sequences in a target nucleic acid. See, e.g., PCT patent publication Nos. WO 89/10977 and 89/11548. In other formats, the detection reagents are supplied in solution.

The development of arraying technologies such as photolithographic synthesis of a nucleic acid array and high density spotting of cDNA products has provided methods for making very large arrays of oligonucleotide probes in very small areas. See U.S. Pat. No. 5,143,854 and PCT patent publication Nos. WO 90/15070 and 92/10092. Microfabricated arrays of large numbers of oligonucleotide probes, called "DNA chips" offer great promise for a wide variety of applications. Such arrays may contain, for example, thousands or millions of probes. Probes may

be formed from, for example, cDNA clones, PCR products, or oligonucleotides and can be used in solution or tethered to a support such as a planar surface (chip) or bead format.

The present invention provides detection kits and reagents, such as nucleic acid arrays, that are based on the novel transcript/exon sequences of the human genome provided herein, particularly the novel transcripts and novel information concerning exon structure of each transcript provided in the Sequence Listing and in Table 1.

Medical Importance of Variable Gene Expression

Variable gene expression, such as alternative splicing (also referred to by such terms as alternate splicing or differential splicing) and alternative start/termination sites, is a fundamentally important mechanism of gene regulation. Alternative splicing refers to the formation of two or more different mature mRNA splice forms from a single gene or pre-mRNA, depending on the combination of exons that are spliced together. Alternative splicing therefore serves as an important means of generating additional protein diversity from the structural information encoded by genes. Furthermore, expression of particular splice forms may differ between, for example, cells, tissues, developmental stages/ages, populations or sexes, and may be altered in certain disease states, such as cancer. Alternative splicing may have a detrimental effect on intercellular interactions and the interaction of various polypeptides and cytokines and thereby lead to diseases such as cancer.

Detection reagents, such as nucleic acid arrays and other multi transcript detection reagent/kit, that utilize detection elements comprised of individual transcripts or exons are capable of detecting alternative splice forms of genes that may be missed by detection reagents that detect only one transcript form. Detection reagents that detect disease-specific splice forms of a gene are useful for disease diagnosis. For example, one or more detection reagents to each exon can be used to determine if an exon is present in a sample and/or detection reagents that span exon/exon boundaries can be used to see if a particular exon/exon splice junction is present and also selects against cross reactivity with genomic DNA.

Alternative splicing plays an important role in a variety of proteins and disease pathways, as the following examples illustrate. Elastin is a protein that is important for providing the elastic properties of the lungs, large blood vessels, and skin. The primary elastin transcript undergoes substantial alternative splicing, and it has been suggested that such alternative splicing of elastin may be population-specific and contribute to aging and pathological conditions in the cardiovascular and pulmonary systems (Indik *et al., Am J Med Genet* 1989 Sep;34(1):81-90).

Nitric oxide proteins are important in numerous physiological processes, such as neurotransmission and muscle relaxation. At least six different isoforms of neuronal nitric oxide mRNA have been identified and found to differ in enzymatic properties. Alternative splicing provides a mechanism to generate this diversity. Furthermore, it has been observed that an alternatively spliced form of neuronal nitric oxide that lacks exon 2 is expressed in many human brain tumors (Brenman *et al., Dev Neurosci* 1997;19(3):224-31).

Alternative splicing of the amyloid precursor protein mRNA, particularly variant splicing of exons 7 and 15, may be involved in the development of Alzheimer's disease (Beyreuther *et al., Ann N Y Acad Sci* 1993 Sep 24;695:91-102).

A number of different estrogen receptor mRNA variants, many of which are generated by alternative splicing, have been identified in breast cancer tissue and may be associated with the development and progression of breast cancer (McGuire *et al., Mol Endocrinol* 1991 Nov;5(11):1571-7).

CD44 is a large family of transmembrane glycoprotein isoforms that are generated from a single gene by alternative splicing and are involved in a variety of cancers. For example, some CD44 isoforms have been found to be causally involved in lung metastasis formation. Furthermore, the expression levels of particular CD44 isoforms are indicative of prognosis in numerous cancers, such as non-Hodgkin lymphomas; gastric, colon, renal, and mammary carcinomas; and in neuroblastomas (Gunthert *et al., Cancer Surv* 1995;24:19-42 and Ponta *et al., Invasion Metastasis* 1994-95;14(1-6):82-6). Therefore, detecting the expression of CD44 alternative splice forms is useful for diagnosing diseases such as these cancers.

Alternative splicing at three positions on the primary fibronectin transcript generates multiple fibronectin polypeptide variants. Furthermore, these different fibronectin variants play specific roles in fibronectin dimer secretion, blood clotting, adhesion to lymphoid cells, skin wound healing, atherosclerosis, and liver fibrosis (Kornblihtt *et al., FASEB J* 1996 Feb;10(2):248-57).

Alternative splicing is important in the differentiation, maintenance, and function of the red blood cell membrane. This is highlighted by the finding that hereditary hemolytic anemias result from mutations that cause defective splicing (Benz *et al., Trans Am Clin Climatol Assoc* 1996;108:78-95).

Platelet derived growth factor (PDGF), which is associated with several diseases including atherosclerosis and neoplasia, undergoes alternative splicing that could affect the function of PDGF (Khachigian *et al., Pathology* 1992 Oct;24(4):280-90); consequently,

alternative splicing of PDGF may play a significant role in diseases such as atherosclerosis and neoplasia.

As an example of the importance of alternative splicing in development, it is well known in the art that sex-specific alternative splicing in *Drosophila* plays an important role in sex determination.

Additionally, exonic splicing enhancers (ESEs) are sequence elements within exons that promote splicing and it has been suggested that many human diseases linked to mutations or polymorphisms within exons may be caused by the inactivation of ESEs, thereby leading to defective splicing (Blencowe, *Trends Biochem Sci* 2000 Mar;25(3):106-10).

As these examples illustrate, such fields as therapeutic/pharmaceutical drug development and disease diagnosis/treatment would greatly benefit from detection kits and reagents that improve the detection of variable gene expression, such as the detection of alternative splice forms.

Using Transcripts/Exons as Detection Elements to Monitor Variable Gene Expression

The transcript sequences and the corresponding exon structure of the transcript disclosed herein are useful in themselves as probe/primer sequences and in the design of such detection element, such as nucleic acid arrays or other detection kits. Transcript sequences with exon structure is particularly useful for studying variable forms of gene expression, such as the expression of alternative splice forms and alternative start/termination sites. As the above examples illustrate, alternative splice forms play important roles in a variety of disease conditions, such as cancer. The importance of detecting alternative splice forms is further highlighted by the finding that nearly 40% of human genes are alternatively spliced (Brett *et al.*, 2000, *FEBS Lett.,* 474, 83); therefore, 40% of all human genes may express alternative transcript forms that are undetectable by conventional detection reagents that are not capable of detecting alternative splice forms of expressed genes.

Individual exons are capable of detecting alternative splice forms that comprise that particular exon, regardless of the combination in which that exon is spliced together with other exons to form an alternative splice form. Therefore, one or more detection elements directed to each single exon can be used to detect any splice form that includes that particular exon. For example, if exon 2 of a six exon gene is used as a detection element (for example, as a probe in a nucleic acid array), that detection element can detect the mRNA splice form of exon 2 with exons 3 and 4, as well as the alternative mRNA splice form of exon 2 with exons 1, 5, and 6. These two different splice forms may have distinct functional properties and one of the two

splice forms may cause a disease condition, or be diagnostic of a disease condition. Exon-based detection elements, such as nucleic acid array probes, may be formed, for example, from exons directly amplified from genomic DNA or synthesized using the sequences provided herein as a reference.

Alternatively, sequences that span an exon/exon junction (see Table 2) can be used to generate detection reagents that are useful in detecting expression and/or splice formation. Such reagents are particularly useful in that detection signal caused by genomic contamination in the sample is greatly reduced.

In addition to alternative splicing, variable gene expression also includes alternative start and termination sites. As with alternative splicing, detection reagents that employ individual exons are useful for detecting transcripts with alternative start and/or alternative termination sites, so long as the transcript includes the exon that comprises the detection element.

Commonly used detection techniques that utilize one transcript form comprised of multiple exons spliced together in a particular combination, such as probes formed from cDNA libraries, are limited in that they will not detect transcripts that are comprised of exons spliced together in a different combination, even if some of the exons are the same. Furthermore, such detection elements may not detect transcripts that comprise alternative start and/or termination sites. This prevents the detection of particular splice forms that may play important roles in, for example, certain disease pathways. Therefore, in certain applications, exon sequences are preferable to larger transcript sequences.

Accordingly, a definite need exists in the art for exons of the human genome provided in a useful form, such as in the form of detection elements of a nucleic acid array or other detection reagent/kit. Exons provided in such a form would be extremely valuable for detecting alternative splice forms and other forms of variable gene expression.

Using Sequences that Span Exon-Exon Junctions as Detection Elements

Sequences that span exon-exon junctions in a transcript are especially useful as detection elements, such as probes in a nucleic acid array. In particular, sequences that span exon-exon junctions eliminate false signals caused by genomic contamination. This is because a detection element comprising two neighboring exons as one contiguous sequence will not hybridize to genomic DNA comprising intervening intronic DNA. Such detection elements will only hybridize to expressed mRNA transcripts in which the exons are connected and the intronic sequence has been removed, thereby forming one contiguous stretch of sequence corresponding to the sequence of the detection element that spans the exon-exon junction.

Exon-exon junctions are provided by the present invention and identified in Table 1. Sequences spanning exon-exon junctions can readily be determined using the exon coordinates provided in Table 1 along with the transcript sequences provided in the Sequence Listing. These detection reagents alone, or in combination with intra-exon probes, can be used to elucidate the splicing and expression pattern of genes within a variety of tissues and/or treatment protocols.

Using Transcripts as Detection Elements to Monitor Gene Expression

Transcript sequences of the human genome are also useful for monitoring gene expression patterns and, in certain circumstances, may be preferable to individual exons for use as detection elements for detecting gene expression. For example, the use of transcripts may be preferred when the goal is to monitor expression of a particular transcript, or group of transcripts, to the exclusion of all other transcripts, such as alternative splice forms. In this situation, using transcripts as detection elements, rather than individual exons, increases specificity and decreases undesired cross hybridization of the detection elements with alternative splice forms.

Accordingly, a definite need exists in the art for transcripts of the human genome, as well as exons, provided in a useful form, such as in the form of detection elements in a detection reagent/kit, such as in a nucleic acid array. Transcripts provided in such a form are useful for monitoring particular forms of gene expression with a high degree of specificity. Such detection elements can readily be generated using the sequence information provided herein.

SUMMARY OF THE INVENTION

The present invention is based on the sequencing and assembly of the human genome. The present invention provides the primary nucleotide sequence of the coding portions of the human genome in a series of predicted transcript sequences generated from the assembled and annotated human genome (SEQ ID NOS:1-39010). Furthermore, the position of each exon contained within these transcripts is identified in Table 1. Individual exon sequences can readily be determined using the transcript sequences of SEQ ID NOS:1-39010 along with the coordinates of each exon within it's respective transcript, as provided in Table 1. This information can be used to readily generate nucleic acid detection reagents and kits, such as nucleic acid arrays. In particular, detection reagents are provided that comprise at least one detection element, wherein at least one detection element comprises a transcript selected from SEQ ID NOS:1-39010. In preferred embodiments, at least one detection element of the detection reagent comprises an exon identified in Table 1. In other preferred embodiments, the detection reagent is a nucleic acid array and the detection elements may be, for example, probes attached

to the surface of the array. Furthermore, in other preferred embodiments, the detection reagent comprises 10,000 or more detection elements, one or more from each of the novel transcripts/exons disclosed herein.

Detection elements that comprise a transcript sequence or an exon, particularly an exon selected from Table 1, allow one to identify variable forms of gene expression, such as different splice forms of genes containing the exon of the detection element. Variable forms of gene expression, such as alternative splicing, may have important tissue-specific, disease-specific, or development-specific expression patterns. Such variable forms of gene expression may go undetected by conventional detection techniques used in gene expression studies. Detection elements that comprise a transcript, particularly a transcript selected from SEQ ID NOS:1-39010, allow one to monitor the expression of the transcript that comprises the detection element with a high degree of specificity.

Furthermore, a preferred class of detection elements provided by the present invention comprises sequences spanning exon-exon junctions. Preferred sequences span one exon-exon junction, however, sequences may span any number of exon junctions. Sequences that span exon-exon junctions are particularly useful in that they eliminate false signals caused by genomic contamination. Exon-exon junctions are provided by the present invention and identified in Table 1. Sequences spanning exon-exon junctions can readily be determined using the exon coordinates provided in Table 1 along with the transcript sequences provided in the Sequence Listing.

The present invention provides the nucleotide sequences of the coding portion of the human genome, namely predicted transcript sequences and corresponding exon information, in a form that can be used, analyzed, and commercialized for other uses in addition to detection kits and reagents. For example, the present invention provides the nucleic acid sequences as contiguous strings of primary sequences in a form readable by computers, such as recorded on computer readable media, e.g., magnetic storage media, such as floppy discs, hard disc storage medium, and magnetic tape; optical storage media such as CD-ROM; electrical storage media such as RAM and ROM; and hybrids of these categories such as magnetic/optical storage media. The present invention specifically provides a CD-R that comprises this sequence information (in the form of a Sequence Listing, provided in file SEQLIST.TXT on the accompanying CD labeled CL001101CDA). Such compositions are useful for, for example, for virtual northern blot analysis, BLAST searching, discovery and validation of drug targets, and for comparative genomic studies between genomes of different organisms.

The present invention further provides systems, particularly computer-based systems, which contain the primary sequence information of the present invention stored in data storage means. Such systems are designed to identify commercially important fragments of the human genome.

Another embodiment of the present invention is directed to isolated fragments, and collections of fragments, of the human genome. The fragments of the human genome include peptide-coding fragments, such as transcripts and exons. The transcript sequences (SEQ ID NOS:1-39010) are provided in the Sequence Listing, which is provided in file SEQLIST.TXT, and the exon elements that each transcript is comprised of are provided in Table 1, which is provided in file TABLE1.TXT. Both files are provided on the accompanying CD labeled CL001101CDA.

As discussed above, the present invention includes detection reagents and kits, such as nucleic acid arrays and microfluidic devices, that comprise one or more fragments of the human genome of the present invention, particularly transcript sequences and/or isolated exon sequences. The kits, such as arrays, can be used to track the expression of many exons or genes, even all of the exons or genes, or rationally selected subsets thereof, contained in the human genome.

The identification of the coding set of sequences from the human genome will be of great value for a variety of commercial purposes. Many fragments of the human genome will be immediately characterized by similarity searches against protein and nucleic acid databases and by identifying structural motifs present in protein domains and will be of immediate value to researchers and for the production of proteins or to control gene expression. A specific example concerns secreted proteins, ion channels and G-protein coupled receptors. The biological significance of secreted proteins for controlling cell signaling, differentiation and proliferation is well known.

Further, the development of therapeutic proteins and protein targets for human intervention typically involves identifying a protein that can serve as a target for the development of a small molecule modulator. Many classes of proteins are well characterized as suitable pharmaceutical drugs (protein therapeutics or modified forms thereof) and/or drug targets. These include, but are not limited to, secreted proteins, GPCRs and ion channels.

Brief Description of the Files contained on CD labeled CL001101CDA

1) File SEQLIST.TXT provides the Sequence Listing of the transcript sequences of the present invention in text (ASCII) format. The file size is 50.7 MB.

2) File TABLE1.TXT provides Table 1, which gives detailed information on exon structure for each of the transcript sequences in the Sequence Listing. The size of this file is 15.2 MB and is stored in text (ASCII) format.

Brief Description of Table 1

Table 1 gives the results of detailed computer analysis of the human genome. Table 1 provides information on every identified human transcript and exon comprising every gene/coding region of the human genome, as follows:

The SEQ ID NO: of each transcript sequence (corresponding to SEQ ID NOS:1-39010 provided in the file, SEQLIST.TXT), a Celera UID identifying number for each transcript, a Celera CT identifying number for each transcript, numbers corresponding to each predicted exon contained within each transcript, predicted exon boundaries (indicating exon-exon junctions) identified by coordinates within the corresponding transcript, and supporting evidence for the existence of each exon and/or transcript, where available (H = human EST/cDNA support, R = rodent EST/cDNA support, M = mouse genomic support, and P = protein homology).

Brief Description of the Figure

The figure provides a block diagram of a computer system 102 that can be used to implement the computer-based systems of the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

The present invention is based on the sequencing and assembly of the human genome. In this process, the primary nucleotide sequence of over 30 million nucleic acid fragments, from about 400 to about 600 nucleotides in length, was determined. These fragments were assembled using the Celera Assembler. After assembly, the sequences were analyzed with various computer packages and compared with all external data sources. The result of this analysis was the identification of 39010 predicted protein-coding transcripts contained in the human genome. The present invention provides the nucleic acid sequences of these transcripts (SEQ ID NOS:1-39010), along with corresponding exon information, in a form that can be used, for example, to readily develop nucleic acid detection kits and reagents, such as nucleic acid arrays.

The present invention provides the nucleotide sequences of the coding sequences of the human genome, including transcript sequences (SEQ ID NOS:1-39010) and corresponding exon information (provided in Table 1), in a form that can be readily used, analyzed, and interpreted

by a skilled artisan. In one embodiment, the sequences are provided as contiguous strings of primary sequence information corresponding to the nucleotide sequences provided in SEQ ID NOS:1-39010, and/or the exons identified in Table 1; the delineated nucleotide sequence of each exon can readily be determined using the transcript sequences of SEQ ID NOS:1-39010 along with the coordinates of each exon within it's respective transcript, as provided in Table 1. The exon information is provided in file TABLE1.TXT and the transcript sequences are provided in file SEQLIST.txt; both of these files are provided on the accompanying CD labeled CL001101CDA. The information in these files has many commercially important uses. For example, the transcript/exon sequences and structural information provided herein can be used to generate commercially valuable nucleic acid or peptide fragments, design and develop probes/primers, and to develop detection reagents and kits such as gene expression arrays. Furthermore, the sequence and structural information provided herein is valuable for a wide variety of commercially important computer-based biological analysis, such as virtual northern blot analysis of gene expression, BLAST searching, or comparative genomic analysis of different organisms. Uses such as these enable the identification and validation of commercially important genes and gene products, as well as diagnostic kits, therapeutics agents, and drug targets.

In other embodiments, the sequences of the present invention are represented by a detection reagent/kit that is capable of identifying mRNA sequences that hybridize to any particular exonic or transcript sequence provided herein. In particular, detection reagents and kits are provided that comprise at least one detection element, wherein at least one detection element comprises a transcript selected from SEQ ID NOS:1-39010, or a portion thereof. In preferred embodiments, at least one detection element of the detection reagent comprises an exon specific detection element identified in Table 1. In other preferred embodiments, at least one detection element of the detection reagent spans at least one exon-exon junction; exon-exon junctions are identified in Table 1. Furthermore, in preferred embodiments, the detection reagent/kit is a nucleic acid array and the detection elements may be, for example, probes attached to the surface of the array. Other preferred detection reagents/detection elements include TaqMan probe/primer sets, for monitoring gene expression using the TaqMan 5' nuclease PCR assay. Furthermore, in most of the preferred embodiments, the detection reagent comprises about more than one detection element (sequence) and preferably, 10,000 or more such detection elements. Such detection reagents can be used to track the expression of many genes/transcripts, or transcript processing, even all of the transcripts/genes/exons, or rationally selected subsets thereof, contained in the human genome.

As used herein, "detection elements" correspond to an element, such as a nucleic acid probe, a probe/primer pair, or a binding aptamer, that is capable of selectively binding a transcript or exon sequence provided by the present invention, or a fragment thereof. Such detection elements include, for example, isolated oligonucleotides comprising the transcript/exon sequences provided herein, provided in a format such as in an array or in a TaqMan 5' nuclease PCR assay. Detection elements, such as probes/primers, may be, for example, attached to a solid support (e.g., in arrays) or supplied in solution (e.g., probe/primer sets for enzymatic reactions such as PCR or RT-PCR).

Additionally, "detection elements" also include the transcript/exon sequences and/or structural information provided herein implemented in a computer-based system. For example, the transcript/exon sequences provided herein may be used as detection elements for searching a computer-based database of sequence or expression information, such as for sequence similarity searching, virtual northern blot analysis, BLAST searching, gene discovery/validation, gene functional analysis, or comparative genomic/expression studies between different individuals, species/organisms, or disease conditions.

Furthermore, one of the preferred classes of detection elements provided by the present invention comprises detection elements that span an exon-exon junction in a transcript. Preferred detection elements span one exon-exon junction. However, detection elements may span any number of exon-exon junctions within a transcript. Detection elements that span exon-exon junctions are particularly useful in that they eliminate false signals caused by genomic contamination. Exon-exon junctions are identified in Table 1. Sequences spanning exon-exon junctions can readily be determined using the exon coordinates provided in Table 1 along with the transcript sequences provided in the Sequence Listing. Thus, references herein to exon, transcript, or gene sequences also include sequences spanning one or more exon-exon junctions.

"Detection reagents" and "detection kits" refer to any system or technology platform that utilizes detection elements comprising nucleic acid or peptide sequences/molecules/fragments corresponding to the transcripts/exons of the present invention, as described above. Thus, detection reagents or detection kits may refer to, for example, nucleic acid arrays (which may also be referred to by such terms as "DNA chips", "biochips", or "microarrays"), the TaqMan 5' nuclease PCR assay system and probe/primer sets, or other enzymatic or PCR-based assay systems, solutions of probes and/or primers, compartmentalized kits, dot-blot or reverse dot-blot systems, sequencing systems, microfluidic systems, mass spec systems, and various computer-based systems such as databases of nucleic acid sequences, protein sequences, or expressed sequences.

The term "transcript" is generally used herein to refer to coding or expressed segments of the human genome that comprise a set of one or more exons that form a mature mRNA molecule upon transcription/expression. The term "transcript" is also used herein to refer to the mRNA transcript molecule, as well as the set of exons in genomic DNA that comprise the mRNA transcript molecule. "Transcripts" may also be referred to herein as "genes", and vice versa, in order to refer to coding portions of genes or open reading frames (ORFs) that correspond to the transcript/exon sequences provided herein.

As used herein, a "representative fragment of the nucleotide sequence provided herein" refers to any portion of these sequences that are not presently represented within a publicly available database, or more particularly to a collection of fragments, where at least one of the members of the collection is unknown, or the entire set has never been described in it's entirety.

Those in the art will readily recognize that detection elements that are comprised of nucleic acid molecules may be supplied as double stranded molecules and that reference to a particular sequence on one strand refers, as well, to the corresponding complementary sequence on the opposite strand. Thus reference to an adenine, a thymine (uridine), a cytosine, or a guanine on one strand of a nucleic acid molecule is also intended to include the thymine (uridine), adenine, guanine, or cytosine, respectively, at the corresponding sites on a complementary strand of the nucleic acid molecule. Thus, reference may be made to either strand in order to refer to a particular nucleic acid sequence or detection element. Oligonucleotide, such as probes and primers, may be based on, or hybridize to, either strand. Throughout the text, reference is generally made to the protein-coding strand, only for the purpose of convenience.

The nucleotide sequence information provided herein was obtained by sequencing the human genome using a shotgun sequencing method known in the art. The nucleotide sequences provided herein are highly accurate, although not necessarily a 100% perfect, representation of the set of exonic nucleotide sequences of the human genome.

Using the information provided herein together with routine cloning and sequencing methods, one of ordinary skill in the art is able to identify, clone and sequence all "representative fragments" of interest including transcripts/exons encoding a large variety of human proteins. In very rare instances, this may reveal a nucleotide sequence error present in the nucleotide sequence disclosed herein. Thus, once the present invention is made available (i.e., the information in the Sequence Listing and Table 1 in a useable form), resolving a rare sequencing error would be well within the skill of the art. Nucleotide sequence editing software is publicly available.

Even if all of the very rare sequencing errors in the sequences herein disclosed were corrected, the resulting nucleotide sequence would still be at least 90% identical, and more likely 99% identical, and most likely 99.99% identical to the nucleotide sequence provided herein.

Thus, the present invention further provides nucleotide sequences that are at least 90% identical, or greater, to the nucleotide sequences of the present invention in a form that can be readily used, analyzed and interpreted by a skilled artisan. Methods for determining whether a nucleotide sequence is at least 90% identical to the nucleotide sequence of the present invention are routine and readily available to a skilled artisan. For example, the well known BLAST algorithm can be used to generate the percent identity of nucleotide sequences.

The present invention also encompasses novel amino acid sequences/proteins/peptides encoded by the transcripts/exons provided herein. Although these encoded amino acid sequences are not explicitly given, such amino acid sequences can readily be determined using the transcript/exon sequences and structural information provided herein in combination with the universal genetic code. Amino acid sequences can be readily generated by numerous algorithms or computer programs commonly used in the art that simply translate the protein-coding nucleic acid sequences provided herein into amino acid sequences based on the universal genetic code. Such amino acid/peptide sequences have commercially valuable uses similar to those described herein for the transcript/exon nucleic acid sequences/fragments of the present invention, such as design of protein detection reagents and computer-based biological analysis, for identification of commercially important proteins.

Nucleic Acid Fragments

Another embodiment of the present invention is directed to isolated fragments of the human genome, particularly those in the form of detection elements or sets of detection elements. The fragments of the human genome of the present invention include, but are not limited to, fragments that encode peptides, particularly genes, exons, and transcripts identified and described in the Sequence Listing (file SEQLIST.TXT) and in Table 1 (file TABLE1.TXT), provided on the accompanying CD labeled CL001101CDA. Such isolated fragments of the human genome, comprising the exon and/or transcript sequences provided herein and fragments thereof, are particularly useful as detection elements, such as for use as probes in a nucleic acid array, for detecting gene expression and other uses.

For example, the nucleic acid molecules/fragments of the present invention, corresponding to the transcript/exon sequences provided herein, are useful as probes, primers, chemical intermediates, and in biological assays for genes of the present invention, particularly

gene expression assays. The probes/primers can correspond to one or more of the exons provided in Table 1, or one or more of the transcripts provided in the Sequence Listing, or may span one or more exon-exon junctions identified in Table 1, or can correspond to a specific region 5' and/or 3' to a transcript or exon provided herein. The transcript/exon sequences and structural information provided herein are also useful for isolating or amplifying any given exon or transcript/gene fragment of the present invention and for designing a variety of gene, or gene expression, detection reagent/kits.

A probe/primer may comprise, for example, a substantially purified exon or transcript molecule or an oligonucleotide or oligonucleotide pair that flanks a defined transcript/exon sequence. A probe/primer comprising an exon or transcript molecule may comprise the full-length exon or transcript sequence, as provided herein, or any portion thereof. A probe/primer comprising an exon or transcript sequence may also include 5' or 3' flanking nucleic acid sequences, depending on the particular assay. Oligonucleotide probes/primers may be shorter molecules that comprise a nucleotide sequence that hybridizes under stringent conditions to at least about 5, 12, 20, 25, 40, 50, 100 or more consecutive nucleotides that comprise a unique sequence specific to the target exon or transcript/gene. Depending on the particular application, the consecutive nucleotides can either include the target exon or transcript, or be a specific region in close enough proximity 5' and/or 3' to the exon or transcript to carry out the desired assay.

Furthermore, a preferred class of nucleic acid fragments are those that span exon-exon junctions. Preferred fragments span one exon-exon junction. However, fragments may span any number of exon-exon junctions within a transcript. Nucleic acid fragments that span exon-exon junctions are particularly useful, when used as detection elements such as probes in an array, in that they eliminate false signals caused by genomic contamination. Exon-exon junctions are identified in Table 1. Nucleic acid fragments spanning exon-exon junctions can readily be determined using the exon coordinates provided in Table 1 along with the transcript sequences provided in the Sequence Listing.

The isolated nucleic acid molecules of the present invention include, but are not limited to, double-stranded or single-stranded DNA or RNA, such as mRNA, cDNA, or genomic DNA comprising the exons or transcript sequences provided herein. Isolated nucleic acid molecules may be obtained, for example, by cloning or PCR amplification, or produced by chemical synthetic techniques or by a combination thereof. Single-stranded nucleic acid can be the coding strand (sense strand) or the non-coding strand (anti-sense strand). Double-stranded RNA molecules are useful for, for example, RNA interference, or gene silencing, which can be used to

turn genes off in order to elucidate their function and may be useful therapeutic agents for turning off defective, disease-causing genes (see Plasterk *et al., Curr Opin Genet Dev* 2000 Oct;10(5):562-7; Bosher *et al., Nat Cell Biol* 2000 Feb;2(2):E31-6; and Hunter, *Curr Biol* 1999 Jun 17;9(12):R440-2).

"Nucleotide sequence" may refer to either a heteropolymer of deoxyribonucleotides, in the case of DNA, or a heteropolymer of ribonucleotides, in the case of RNA. DNA or RNA segments may be assembled, for example, from fragments of the human genome or single nucleotides, short oligonucleotide linkers, or from a series of oligonucleotides, to provide a synthetic nucleic acid molecule.

The present invention provides isolated nucleic acid molecules that contain one or more exons or transcripts disclosed by the present invention. Such nucleic acid molecules will consist of, consist essentially of, or comprise one or more exons or transcripts of the present invention. The nucleic acid molecule can have additional nucleic acid residues, such as nucleic acid residues that are naturally associated with it or heterologous nucleotide sequences.

As used herein, an "isolated" nucleic acid molecule is one that contains an exon and/or transcript of the present invention and is separated from other nucleic acid present in the natural source of the nucleic acid. The isolated nucleic acid, as used herein, will be comprised of one or more exons and/or transcripts disclosed by the present invention. The isolated nucleic acid may have flanking nucleotide sequence on either side of the exon or transcript depending on the particular use of the isolated nucleic acid or assay involved. The flanking sequence may be, for example, up to about 5,000 bases; 2,500 bases; 1,000 bases; 500 bases; 100 bases, 50 bases, 30 bases, 20 bases, or 10 bases on either side of an exon or transcript, for detection reagents. The important point is that the nucleic acid is isolated from remote and unimportant flanking sequences and is of appropriate length such that it can be subjected to the specific manipulations or uses such as recombinant expression, preparation of probes and primers for expression analysis, and other uses specific to the transcript/exon sequences.

As used herein, an "isolated nucleic acid molecule" or an "isolated fragment of the human genome" refers to a nucleic acid molecule possessing a specific nucleotide sequence which has been subjected to purification means to reduce, from the composition, the number of compounds which are normally associated with the composition. A variety of purification means that are well known in the art can be used to generate the isolated fragments of the present invention. These include, but are not limited to, methods that separate constituents of a solution based on charge, solubility, or size. Moreover, an "isolated" nucleic acid molecule, such as an mRNA molecule containing a transcript sequence of the present invention or an exon isolated from

genomic DNA, can be substantially free of other cellular material, or culture medium when produced by recombinant techniques, or chemical precursors or other chemicals when chemically synthesized. However, the nucleic acid molecule can be fused to other coding or regulatory sequences and still be considered isolated. For example, recombinant DNA molecules contained in a vector are considered isolated. Further examples of isolated DNA molecules include recombinant DNA molecules maintained in heterologous host cells or purified (partially or substantially) DNA molecules in solution. Isolated RNA molecules include in vivo or in vitro RNA transcripts comprising the sequences of the present invention. Isolated nucleic acid molecules according to the present invention further include such molecules produced synthetically.

In one embodiment, human DNA can be mechanically sheared to produce fragments of about 2kb, 10kb, or 15-20 kb in length. These fragments can then be used to generate a human library by inserting them into plasmid vectors (or lambda vectors) using methods well known in the art. Primers flanking, for example a gene or exon, can then be generated using nucleotide sequence information provided in the present invention. PCR cloning can then be used to isolate the gene or exon from the human DNA library. PCR cloning is well known in the art. Thus, given the availability of the present identified gene coding sequences of the human genome, it is routine experimentation to isolate any gene or exon, or fragments thereof, particularly using the information provided in file, TABLE1.TXT, provided on the accompanying CD labeled CL001101CDA. Particularly useful is the generation of nucleic acid fragments comprising one or more exons of a gene, particularly those identified herein. Such fragments can be applied to an array, microfluidic device or other detection kit format and used to detect expression of a gene (see below).

The sequences falling within the scope of the present invention are not limited to the specific sequences herein described, but also include allelic and species variations thereof. Allelic and species variations can be routinely determined by comparing the sequences provided by the present invention, or a representative fragment thereof, with sequences from other isolates from the same species (allelic variations) or from other species (species variations). Sequence comparisons with other nucleic acid isolates to determine allelic or species variation can be readily accomplished using the transcript/exon sequences and structural information provided herein. For example, primers for re-sequencing any particular transcript, exon, or fragment thereof can be readily designed based on the sequences provided herein. Such re-sequencing is useful for detecting polymorphisms, such as SNPs, in the transcripts/exons provided herein. Furthermore, such SNPs, being in protein coding regions, are of significant commercial value

since they may change the encoded protein sequence and thereby play a direct role in disease development and progression. Such SNPs are important targets for therapeutic/drug development, and may also serve as important diagnostic/prognostic markers. Thus, the transcript/exon sequences and structural information provided herein is a commercially valuable resource for SNP detection.

To accommodate codon variability, the present invention also encompasses nucleic acid molecules coding for the same amino acid sequences as do the specific transcript/exon sequences disclosed herein. In other words, in the transcript/exon sequences disclosed herein, substitution of one codon for another that encodes the same amino acid is expressly contemplated.

The present invention further provides related nucleic acid molecules that hybridize under stringent conditions to the nucleic acid molecules disclosed herein. As used herein, the term "hybridizes under stringent conditions" is intended to describe conditions for hybridization and washing under which nucleotide sequences encoding a peptide at least 60-70% homologous to each other typically remain hybridized to each other. The conditions can be such that sequences at least about 60%, at least about 70%, or at least about 80%, or at least about 90% or more homologous to each other typically remain hybridized to each other. Such stringent conditions are known to those skilled in the art and can be found in Current Protocols in Molecular Biology, John Wiley & Sons, N.Y. (1989), 6.3.1-6.3.6. One example of stringent hybridization conditions are hybridization in 6X sodium chloride/sodium citrate (SSC) at about 45°C, followed by one or more washes in 0.2 X SSC, 0.1% SDS at 50-65°C. Examples of moderate to low stringency hybridization conditions are well known in the art.

Any specific sequence disclosed herein can be readily screened for errors by resequencing a particular fragment, such as an exon or transcript, in both directions (i.e., sequence both strands). Alternatively, error screening can be performed by sequencing corresponding polynucleotides of human origin isolated by using part or all of the fragments in question as a probe or primer.

Each of the transcripts/exons of the human genome, including sequences and isolated nucleic acid molecules, can be routinely characterized using the computer system of the present invention and can be used in numerous ways as polynucleotide reagents. For example, isolated nucleic acid molecules comprising at least one of the exon or transcript sequences provided herein, can be used as diagnostic probes or diagnostic amplification primers to detect the expression of a particular exon, exon set, gene, or gene set. This is particularly useful in the form of nucleic acid arrays wherein 100 or more, 1000 or more, 5000 or more, 10,000 or more,

or even most to all of the exons/transcripts provided by the present invention are implemented in a single array.

Nucleic Acid Arrays and Detection Kits and Reagents

The present invention provides detection kits and reagents, such as, but not limited to, arrays, TaqMan probe/primer sets, and various compartmentalized kits, comprising detection elements, such as nucleic acid probes, that are based on the sequence information provided by the present invention, particularly the transcript sequences (SEQ ID NOS:1-39010) or exon sequences (exon information is provided in Table 1).

As used herein "Arrays" or "Microarrays" refers to an array of distinct polynucleotides or oligonucleotides synthesized on a substrate, such as paper, nylon or other type of membrane, filter, chip, glass slide, plastic, silicon, gold, gel or any other suitable solid, or semi-solid support. Arrays may also be based on fiber-optics and comprise, for example, probes attached to beads at the ends of fiber-optic bundles (see Walt, *Science* 287, 451 (2000), Michael *et al., Anal. Chem* 70, 1242-1248 (1998), and Ferguson *et al., Nature Biotechnology* 14, 1681-1684 (1996)). In one embodiment, the microarray is prepared and used according to the methods described in US Patent 5,837,832 (Chee et al.), PCT application WO95/11995 (Chee et al.), Lockhart, D. J. et al. (1996; *Nat. Biotech.* 14: 1675-1680) and Schena, M. et al. (1996; *Proc. Natl. Acad. Sci.* 93: 10614-10619), all of which are incorporated herein in their entirety by reference. In other embodiments, such arrays are produced by the methods described by Brown et. al., US Patent No. 5,807,522. Hybridization and scanning of arrays is also described in PCT application WO 92/10092 and EP785280. The use of microarrays of oligonucleotides or polynucleotides for capturing complementary polynucleotides from expressed genes is also described in Schena *et al, Science*, 270: 467-469 (1995); DeRisi *et al, Science*, 278: 680-686 (1997); Chee *et al, Science*, 274: 610-614 (1996). Additionally, Freeman *et al.* (*Biotechniques* 29, 1042-1055 (2000), Lockhart *et al.* (*Nature* 405, 827-836 (2000)), and Zweiger (*Trends in Biotechnology* 17, 429-436 (1999)) provide reviews of nucleic acid arrays for gene expression analysis and other uses; also see *Nature Genetics* 21 (Suppl.), 1-60 (1999) and Meldrum, *Genome Research*, 10:1288-1303 (2000) for an overview of array technology.

For example, gene expression kits and reagents, such as arrays or sets of probe containing beads, may contain one or more detection elements, such as oligonucleotide probes or pairs of probes, that hybridize at or near each exon or gene corresponding to the exon/transcript sequences provided by the present invention. A plurality of oligonucleotide probes may be included in the kit to simultaneously assay large numbers of genes/exons, at least one of which is

one of the genes/exons of the present invention and novel to the present disclosure. In some kits, such as arrays, the oligonucleotide probes are provided immobilized to a substrate. For example, the same substrate can comprise oligonucleotide probes for detecting at least 1; 10; 100; 1000; 10,000 or most or substantially all of the genes/transcripts or exons provided by the present invention. Any number of probes, or other detection elements, may be utilized in a detection reagent, depending on the particular technology platform and objective. For example, a typical array may contain hundreds, thousands to millions of individual synthetic DNA probes arranged in a grid-like pattern and miniaturized to the size of a dime, each corresponding to a particular exon or transcript/gene. Preferably, probes are attached to a solid support in an ordered, addressable array. Customized arrays that utilize the exon and/or gene/transcript sequences provided by the present invention can be produced by various manufacturers. For example, arrays with over 250,000 oligonucleotide probes or 10,000 cDNAs per square centimeter are readily available (see Lipshutz *et al., Nature Genetics*, 21, 20-24 (1999) and Bowtell *et al., Nature Genetics*, 21, 25-32 (1999). In some arrays, electric fields can be applied to the array to speed hybridization reactions (see Edman *et al., Nucleic Acids Res.* 25, 4907-4914 (1997) and Sosnowski *et al., Proc. Natl. Acad. Sci. USA* 94, 1119-1123 (1997)). Arrays have been previously produced for completely sequenced organisms, such as *Saccharomyces cerevisiae*, that comprise probes for every identified gene in the organism's genome (see DeRisi *et al., Science* 278, 680-686 (1997) and Wodicka *et al., Nature Biotechnology* 15, 1359-1367 (1997)).

The microarray or detection kit is preferably composed of a large number of unique nucleic acid sequences, usually either synthetic antisense oligonucleotides or fragments of cDNAs, fixed to a solid support. Probes may comprise either single- or double-stranded nucleic acid molecules. Oligonucleotides may be about 6-60 nucleotides in length, more preferably 15-30 nucleotides in length, and most preferably about 20-25 nucleotides in length. For a certain type of microarray or detection kit, it may be preferable to use oligonucleotides that are only 7-20 nucleotides in length. For others, such as cDNA, longer lengths are possible and preferable. These can be of the order of 1kb-5kb or more in length and can comprise the entire length of a transcript or exon sequence provided herein or can comprise a short fragment of the transcript/exon, such as in exon-exon junction spanning detection elements.

The microarray or detection kit may contain oligonucleotides that cover, for example, sequential oligonucleotides that cover the full-length sequence, or unique oligonucleotides selected from particular areas along the length of the sequence, such as in exon-exon boundaries. Additionally, such as in the case of primers for PCR, it may be desirable for oligonucleotides to bind to regions 5' or 3' of the transcripts/exons provided herein, such as to capture the entire

exon or transcript/gene within the amplicon. Polynucleotides used in the microarray or detection kit may be oligonucleotides that are specific to an exon, exons, gene, or genes of interest.

Thus, the chip may comprise an array comprising at least one probe corresponding to the full-length sequence of at least one of the exons and/or transcripts provided by the present invention, sequences spanning one or more exon-exon junctions identified in Table 1, sequences complementary thereto, or fragments thereof. Thus, the sequence of at least one probe of the array is selected from the group consisting of those disclosed in SEQ ID NOS:1-39010 and the exons identified in Table 1, sequences spanning one or more exon-exon junctions identified in Table 1, sequences complementary thereto, and fragments thereof.

In order to produce oligonucleotides to a known sequence for a microarray or detection kit, the exon(s) or gene(s) of interest is typically examined using a computer algorithm that starts at the 5' or at the 3' end of the nucleotide sequence. Typical algorithms will then identify oligomers of defined length that are unique to the exon/gene, have a GC content within a range suitable for hybridization, and lack predicted secondary structure that may interfere with hybridization. In certain situations it may be appropriate to use pairs of oligonucleotides on a microarray or detection kit. For example, pairs of oligonucleotides are particularly useful for detecting mismatch hybridization in high-density arrays that use short oligonucleotides, such as 25-mers; such short oligonucleotides are susceptible to mismatch hybridization due to false priming. In this situation, pairs of oligonucleotides with deliberate mismatches are incorporated to determine the level of mismatch hybridization, which can then be subtracted from the true target signal (Lockhart *et al., Nat. Biotechnology* (1996) 14:1675-1680 and Wodicka *et al., Nat. Biotechnology* (1997) 15:1359-1366). Pairs of oligonucleotide probes are also useful for detecting polymorphisms, particularly SNPs; in these situations, the oligonucleotide pairs are generally designed to be identical except for one nucleotide that preferably is located at or near the center of the sequence. The second oligonucleotide in the pair (mismatched by one) serves as a control. The oligomers are synthesized at designated areas on a substrate using a light-directed chemical process. The substrate may be paper, nylon or other type of membrane, filter, chip, glass slide or any other suitable solid support.

In another aspect, an oligonucleotide may be synthesized on the surface of the substrate by using a chemical coupling procedure and an ink jet application apparatus, as described in PCT application WO95/251116 (Baldeschweiler et al.) which is incorporated herein in its entirety by reference. In another aspect, a "gridded" array analogous to a dot (or slot) blot may be used to arrange and link cDNA fragments or oligonucleotides to the surface of a substrate using, for example, a vacuum system, thermal, UV, mechanical or chemical bonding procedure. An array,

such as those described above, may be produced by hand or by using available devices (slot blot or dot blot apparatus), materials (any suitable solid support), and machines (including robotic instruments), and may contain 8; 24; 96; 384; 1536; 6144; 10,000 or more oligonucleotides, or any other number which lends itself to the efficient use of commercially available instrumentation.

In other embodiments, the array or detection reagent/kit can be produced by spotting cDNA or other nucleic acid molecules onto the surface of a substrate (see Brown et. al., US Patent No. 5,807,522). In such use, PCR amplification of one or more exons or transcripts from genomic DNA can be used to generate a nucleic acid molecule suitable for deposition onto a substrate.

In yet another embodiment, the detection reagent or kit comprises TaqMan probe/primer sets for carrying out the TaqMan PCR assay, such as for detecting gene expression. The TaqMan assay, also known as the 5' nuclease PCR assay, provides a sensitive and rapid means of detecting gene expression. The TaqMan assay detects the accumulation of a specific amplified product during PCR. The TaqMan assay utilizes an oligonucleotide probe labeled with a fluorescent reporter dye at the 5' end of the probe and a quencher dye at the 3' end of the probe. During the PCR reaction, the 5' nuclease activity of DNA polymerase cleaves the probe, thereby separating the reporter dye and the quencher dye and resulting in increased fluorescence of the reporter. Accumulation of PCR product is detected directly by monitoring the increase in fluorescence of the reporter dye. The 5' nuclease activity of DNA polymerase cleaves the probe between the reporter and the quencher only if the probe hybridizes to the target and is amplified during PCR. The probe is designed to hybridize to a target nucleic acid molecule only if the target sequence is complementary to the probe, i.e., if the target sequence comprises the transcript/exon sequence that is used as a probe.

Preferred TaqMan primer and probe sequences can readily be determined using the nucleic acid information provided herein. A number of computer programs, such as Primer-Express, can be used to readily obtain optimal primer/probe sets. It will be apparent to one of skill in the art that the primers and probes based on the nucleic acid and transcript/exon sequences and structural information provided herein are useful as probes or amplification primers for screening for the transcripts/exons provided by the present invention, such as for monitoring gene expression in particular disease conditions, and can be incorporated into a kit format. In particular, genome-wide TaqMan probe/primer sets are specifically contemplated for monitoring the expression of 10,000 or more, or most or all, human genes, or any subset thereof of interest. Such genome-wide TaqMan probe/primer sets can readily be obtained using the

transcript sequences and transcript/exon structural information provided herein, along with a primer/probe design computer program, such as Primer-Express.

Other detection kits and reagents may be based on blotting techniques such as northern blots (for detecting RNA), southern blots (for detecting DNA), or western blots (for detecting proteins) or beads containing detection elements that are well known in the art. The exons and transcript sequences provided by the present invention are well suited for use as detection probes in such techniques.

Direct sequencing, including cDNA sequencing, can also be used to detect the transcripts and/or exons of the present invention. A variety of automated sequencing procedures can be utilized when performing detection/diagnostic assays ((1995) *Biotechniques* 19:448), including sequencing by mass spectrometry (see, e.g., PCT International Publication No. WO 94/16101; Cohen *et al.*, *Adv. Chromatogr.* 36:127-162 (1996); and Griffin *et al.*, *Appl. Biochem. Biotechnol.* 38:147-159 (1993)).

Various other methods useful for gene expression analysis include, but are not limited to, RT-PCR, nuclease protection, clone hybridization, differential display (Liang *et al.*, *Science* 257, 967-971 (1992)), subtractive hybridization, cDNA fingerprinting (Shimkets *et al.*, *Nature Biotechnology* 17, 798-803 (1999), Ivanova, *Nucleic Acids Research* 23, 2954-2958 (1995), Kato, *Nucleic Acids Research* 23, 3685-3690 (1995), and Bachem *et al.*, *Plant J.* 9, 745-753 (1996)), reporter-gene analysis, two-dimensional (2D) gel electrophoresis, mass spectrometry, and serial analysis of gene expression (SAGE) (Velculescu *et al.*, *Science* 270, 484-487 (1995)).

In order to conduct sample analysis using a microarray or other detection reagent/kit, a typical procedure may be similar to the following. The RNA or DNA from a biological sample is made into hybridization probes. The mRNA is isolated, and cDNA is produced and used as a template to make antisense RNA (aRNA). The aRNA is amplified in the presence of fluorescent nucleotides, and labeled probes are incubated with the microarray or detection kit so that the probe sequences hybridize to complementary oligonucleotides of the microarray or detection kit. Incubation conditions may be adjusted so that hybridization occurs with precise complementary matches or with various degrees of less complementarity. After removal of nonhybridized probes, a scanner is used to determine the levels and patterns of fluorescence. The scanned images are examined to determine degree of complementarity and the relative abundance of each oligonucleotide sequence on the microarray or detection kit. The biological samples may be obtained from any bodily fluids (such as blood, urine, saliva, phlegm, gastric juices, etc.), cultured cells, biopsies, or other tissue preparations. A detection system may be used to measure the absence, presence, and amount of hybridization for all of the distinct sequences

simultaneously. This data may be used for purposes including, but not limited to, large-scale correlation studies on the sequences, expression patterns, mutations, variants, or polymorphisms among samples.

Using such arrays, the present invention provides methods to identify the expression of one or more of the exons or transcripts/genes of the present invention. Such methods may comprise incubating a test sample with an array comprising one or more oligonucleotide probes corresponding to at least one exon or transcript of the present invention and assaying for binding of a nucleic acid from the test sample with one or more of the oligonucleotide probes. Such assays will typically involve arrays comprising most, if not all, of the exons or transcripts in the human genome, or rationally selected subsets thereof. The transcript sequences of the human genome are provided in SEQ ID NOS:1-39010 and the exons that these transcripts are comprised of are provided in Table 1.

Conditions for incubating a nucleic acid molecule with a test sample vary. Incubation conditions depend on the format employed in the assay, the detection methods employed, and the type and nature of the nucleic acid molecule used in the assay. One skilled in the art will recognize that any one of the commonly available hybridization, amplification, or array assay formats can readily be adapted to employ the novel fragments of the human genome disclosed herein. Examples of such assays can be found in Chard, T, *An Introduction to Radioimmunoassay and Related Techniques,* Elsevier Science Publishers, Amsterdam, The Netherlands (1986); Bullock, G. R. *et al., Techniques in Immunocytochemistry,* Academic Press, Orlando, FL Vol. 1 (1982), Vol. 2 (1983), Vol. 3 (1985); Tijssen, P., *Practice and Theory of Enzyme Immunoassays: Laboratory Techniques in Biochemistry and Molecular Biology,* Elsevier Science Publishers, Amsterdam, The Netherlands (1985).

The test samples of the present invention include, but are not limited to, nucleic acid extracts, cells, and protein or membrane extracts from cells, which may be obtained from any bodily fluids (such as blood, urine, saliva, phlegm, gastric juices, etc.), cultured cells, biopsies, or other tissue preparations. The test sample used in the above-described methods will vary based on the assay format, nature of the detection method and the tissues, cells or extracts used as the sample to be assayed. Methods of preparing nucleic acid, protein, or cell extracts are well known in the art and can be readily be adapted in order to obtain a sample that is compatible with the system utilized.

In another embodiment of the present invention, kits are provided which contain the necessary reagents to carry out one or more assays for detecting the exons/transcripts/genes of the present invention, such as for gene expression analysis. Specifically, the invention provides a

compartmentalized kit to receive, in close confinement, one or more containers, comprising: (a) a first container comprising at least one nucleic acid molecule that can bind to a fragment of at least one of the exon or transcript sequences disclosed herein, including exon-exon spanning sequences; and (b) one or more other containers comprising wash reagents and/or reagents capable of detecting presence of a bound nucleic acid. Preferred kits will include detection reagents/arrays/chips/microfluidic devices that are capable of detecting the expression of 1 or more, 10 or more, 100 or more, 500 or more, 1000 or more, 10,000 or more, or most or all of the exons or transcripts identified herein that are expressed in humans. One skilled in the art will readily recognize that the previously unidentified exons/transcripts provided by the present invention can be readily incorporated into one of the established kit formats which are well known in the art, particularly expression arrays.

In detail, a compartmentalized kit includes any kit in which reagents are contained in separate containers. Such containers include small glass containers, plastic containers, strips of plastic, glass or paper, or arraying material such as silica. Such containers allow one to efficiently transfer reagents from one compartment to another compartment such that the samples and reagents are not cross-contaminated, and the agents or solutions of each container can be added in a quantitative fashion from one compartment to another. Such kits may typically include a container which will accept the test sample, a container which contains the nucleic acid probe, containers which contain wash reagents (such as phosphate buffered saline, Tris-buffers, etc.), and containers which contain the reagents used to detect the bound probe. The kit can further comprise reagents for PCR, RT-PCR or other enzymatic reactions, and instructions for using the kit. Such compartmentalized kits include multicomponent integrated systems.

Multicomponent integrated systems may also implement the transcript/exon sequences, including exon-exon spanning sequences, provided by the present invention as detection elements. Multicomponent integrated systems include such systems as microfluidic devices, biomedical micro-electro-mechanical systems (bioMEMS), and "lab-on-a-chip" systems (see, for example, US patents 6,153,073, Dubrow *et al.*, and 6,156,181, Parce *et al.*). Such systems miniaturize and compartmentalize processes such as probe/target hybridization, PCR, and capillary electrophoresis reactions in a single functional device, and may be integrated with nucleic acid arrays. An example of such a technique is disclosed in US patent 5,589,136, which describes the integration of PCR amplification and capillary electrophoresis in chips. Multicomponent integrated systems such as microfluidic, bioMEMs, and lab-on-a-chip systems, generally comprise a pattern of microchannels designed onto a glass, silicon, quartz, or plastic wafer included on a microchip. The movements of the samples are controlled by electric,

electroosmotic, or hydrostatic forces applied across different areas of the microchip to create functional microscopic valves and pumps with no moving parts. Varying the voltage controls the liquid flow at intersections between the micro-machined channels and changes the liquid flow rate for pumping across different sections of the microchip.

Medical- and Pharmaceutical-Related Uses

Detection of gene expression, using the transcripts and/or exons of the present invention, is valuable for such uses as disease diagnosis, monitoring disease progression, determining the effects of various treatments/therapeutics, and individualizing medical treatment or drug therapy based on an individual's gene expression patterns. In particular, uses such as these can be achieved using the detection reagents provided by the present invention, such as nucleic acid arrays that utilize the human exon and/or transcript sequences provided by the present invention as detection elements. Genome-wide expression analysis can be conducted in humans using the exons/transcripts provided herein; genome-wide expression analysis has previously been accomplished in yeast (Holstege *et al., Cell* 95, 717-728 (1998)).

Detection reagents, such as arrays, containing the transcripts/exons of the present invention can also be used to probe genomic DNA for changes in gene copy number or allelic imbalances (see Mei *et al., Genome Res* 2000 Aug;10(8):1126-37, Pollack *et al., Nature Genetics* 23, 41-46 (1999), and Pinkel *et al. Nature Genetics* 20, 207-211 (1998)). Such copy number changes/allelic imbalances may be caused by gene or chromosome deletions or duplications, which may occur in cancerous cells and other disorders. Furthermore, identification of genetic/chromosomal changes such as these may facilitate the identification of specific genes, regulatory/control regions, or other genetic elements that play important roles in the disorder, or indicate that a particular chromosomal region harbors such elements.

The sequences and detection reagents of the present invention may be used to determine whether an individual has a mutation or polymorphism, such as a SNP (single nucleotide polymorphism), affecting the level (i.e., the concentration of mRNA or protein in a sample, etc.) or pattern (i.e., the kinetics of expression, rate of decomposition, stability profile, Km, Vmax, etc.) of gene expression in a particular cell, tissue, bodily fluid, disease state, or developmental stage. Such variations in gene expression can be caused, for example, by a SNP in a gene, or in a regulatory/control region(s), such as a promoter, or other gene(s) that controls or affects the expression of the gene. Such an analysis of gene expression can be conducted by screening for mRNA corresponding to the exons and/or transcripts provided by the present invention. Once changes in gene expression patterns are identified, the nucleic acid sequences provided by the

present invention can be used, for example, to design primers/probes for SNP-detection assays to determine if a SNP is responsible for the variation in gene expression patterns. Such SNP-detection assays include, but are not limited to, direct sequencing, mini-sequencing primer extension, and the TaqMan PCR assay, or any other SNP-detection technique known in the art. Furthermore, SNP-detection assays may utilize nucleic acid arrays, mass spec, or other technology platforms used in the art for SNP-detection. Once a SNP is detected that alters gene expression in a manner that contributes to a pathological condition, therapeutic approaches can be targeted at that SNP and, furthermore, that SNP can serve as a diagnostic/prognostic marker for the disease, and may form the basis of a diagnostic kit for the disease. Furthermore, SNPs in the transcript/exon coding sequences provided herein can readily be determined by comparing the sequences provided herein against corresponding transcript/exon sequences from nucleic acid isolates taken from different individuals, such as by re-sequencing or computer-based sequence database comparison. Additionally, changes in the amino acid/protein sequences caused by such SNPs can readily be determined using the sequences provided herein as a reference and the universal genetic code.

Medical gene expression analysis can include the steps of collecting a sample of cells from a patient, isolating mRNA from the cells of the sample, contacting the mRNA sample with one or more probes, based on the exon and/or transcript sequences provided herein, which specifically hybridize to a region of the isolated mRNA containing a target exon/transcript under conditions such that hybridization of the probe with the exon/transcript occurs, and detecting the presence or absence of hybridization. The presence or absence of hybridization, and therefore of the target exon/transcript, can then be correlated with known gene expression patterns in, for example, normal cells/tissues and in cells/tissues in various disease stages in order to, for example, diagnose a disease, determine disease progression, or determine the effect of a particular drug treatment.

The contribution or association of particular gene expression patterns with disease phenotypes enables the transcripts/exons of the present invention to be used to develop superior diagnostic tests based on gene expression/mRNA markers. Such gene expression-based diagnostic tests are useful for identifying individuals who have a gene expression indicative of a specific disease or disease propensity or individuals whose gene expression patterns indicate that a particular drug treatment or therapeutic approach should be utilized. For example, HER2 and the estrogen receptor genes are known to be expressed at increased levels in cancers, such as breast and ovarian cancer (van de Vijver *et al.* (1988) *New Engl. J. Med.* 319, 1239-1245, Berger *et al.* (1988) *Cancer Res.* 48, 1238-1243, and Petrangeli *et al.* (1994) *J. Steroid Biochem. Mol.*

Biol. 49, 327-331) and determining the expression level of these genes may aid physicians in choosing the most effective treatment (McNeil *et al.* (1999) *J. Natl. Cancer Inst.* 91, 110-112, Leinster *et al.* (1998) *Biochem Soc. Symp.* 63, 185-191, and Revillon *et al.* (1998) *Eur. J. Cancer* 34, 791-808). Such diagnostics may be based on a single transcript/gene or exon, a group of transcripts/genes or exons, or most or all transcripts/genes or exons provided by the present invention.

The invention further provides a method for identifying a compound that can be used to treat a disorder associated with expression of a disease-associated gene or variable, disease-associated, expression of a normal gene. Forms of gene expression such as these are collectively referred to herein as disease-associated gene expression, and may contribute to, for example, disease or developmental disorders. The method typically includes assaying the ability of the compound to modulate the activity and/or expression of the target nucleic acid and thus identifying a compound that can be used to treat a disorder characterized by undesired activity or expression of the nucleic acid.

The assays for disease-associated nucleic acid expression can be accomplished using the transcript and/or exon sequences provided by the present invention as gene expression detection elements, such as probes in a nucleic acid array. The assay for disease-associated nucleic acid expression can involve direct assay of nucleic acid levels, such as mRNA levels, or on collateral compounds involved in the signal pathway. Further, the expression of genes that are up- or down regulated in response to the disease-associated protein signal pathway can also be assayed. In this embodiment the regulatory regions of these genes can be operably linked to a reporter gene such as luciferase.

Thus, modulators of disease-associated gene expression can be identified in a method wherein a cell is contacted with a candidate compound, such as a drug or small molecule, and the expression of mRNA determined. The level of expression of disease-associated mRNA in the presence of the candidate compound is compared to the level of expression of disease-associated mRNA in the absence of the candidate compound. The candidate compound can then be identified as a modulator of nucleic acid expression based on this comparison and be used, for example, to treat a disorder characterized by disease-associated gene expression. When expression of mRNA is statistically significantly greater in the presence of the candidate compound than in its absence, the candidate compound is identified as a stimulator of nucleic acid expression. When nucleic acid expression is statistically significantly less in the presence of the candidate compound than in its absence, the candidate compound is identified as an inhibitor of nucleic acid expression.

The invention further provides methods of treatment, with one or more of the genes/transcripts/exons provided by the present invention as a target, using a compound identified through drug screening using the transcript/exon sequences provided herein, as a gene modulator to modulate nucleic acid expression. Modulation includes both up-regulation (i.e. activation or agonization) or down-regulation (suppression or antagonization) of nucleic acid expression. These methods of treatment include the step of administering the modulators of gene expression in a pharmaceutical composition to a subject in need of such treatment.

The exon/transcript sequences provided herein are also useful for monitoring the effectiveness of modulating compounds on the expression or activity of a gene in clinical trials or in a treatment regimen. Thus, the gene expression pattern can serve as a barometer for the continuing effectiveness of treatment with the compound, particularly with compounds to which a patient can develop resistance. The gene expression pattern can also serve as a marker indicative of a physiological response of the affected cells to the compound. Accordingly, such monitoring would allow either increased administration of the compound or the administration of alternative compounds to which the patient has not become resistant. Similarly, if the level of nucleic acid expression falls below a desirable level, administration of the compound could be commensurately decreased. Therefore, the transcript/exon sequences of the present invention are particularly useful for improving the process of drug development by allowing changes in gene expression patterns in response to candidate compounds/drugs to be determined; such changes in gene expression patterns can be analyzed to determine compound/drug efficacy and/or toxicity. This not only improves the safety of clinical trials, but also will enhance the chances that the trial will demonstrate statistically significant efficacy by allowing the clinical trials to be adjusted in response to different gene expression patterns observed in different patients in response to a candidate compound/drug. Furthermore, gene expression analysis using the transcripts/exons of the present invention may help explain why certain, previously developed drugs performed poorly in clinical trials and may help identify a subset of the population that would benefit from a drug that had previously performed poorly in clinical trials, thereby "rescuing" previously developed drugs.

Gene expression analysis using the detection reagents of the present invention is also useful for determining the target of a drug. For example, gene expression patterns in cells treated with a drug can be compared to gene expression patterns in cells that have had individual genes, particularly genes corresponding to the exons/transcripts provided herein, inactivated. A similar gene expression pattern would indicate that the drug may target the gene that had been inactivated.

Gene expression analysis, using the transcripts/exons provided by the present invention, may also be useful in forensic and medicolegal investigations. For example, post-mortem gene expression analysis may provide clues as to cause of death or time of death, may indicate exposure to toxic compounds or drugs, and may aid in identification.

Examples of other important uses of the transcripts/exons provided herein for gene expression include, but are not limited to, determining the toxicological consequences of altered gene expression (Pennie, *Toxicol Lett* 2000 Mar 15; 112-113: 473-7), understanding changes in gene expression in response to infection (Manger *et al., Curr Opin Immunol* 2000 Apr;12(2):215-8) and modulating gene expression to enhance the immune response, and regulating the expression of genes delivered through gene therapy (Clackson, *Gene Ther* 2000 Jan;7(2):120-5).

Expression Modulating Fragments

The present invention is useful for unraveling and characterizing the complex genetic network involved in the regulation and control of gene expression. For example, the present invention facilitates the identification and characterization of regulatory/control elements in the human genome, referred to herein as "expression modulating fragments" (EMFs), or expression modulating elements/sequences. As used herein, an "expression modulating fragment," means a series of nucleotide molecules that modulate the expression of an operably linked gene/transcripts or another EMF. EMFs may also include gene products such as transcriptional activators and repressors. Sets of co-regulated genes, referred to as "regulons", can also be identified. Genomic features such as novel EMFs and regulons can be identified, for example, through genome-wide expression analysis using arrays comprising the exons/transcripts provided by the present invention. Genomic sequence motifs that are statistically over-abundant in regions close to similarly expressed genes, particularly in 5' regions, may be identified as novel EMFs, such as *cis*-regulatory elements. Furthermore, using genome-wide expression analysis, one can determine whether an EMF has a global effect (affects a large number of genes, or all genes) or a specific effect (affects a small number of genes, or a single gene) (Holstege *et al., Cell* 95, 717-728 (1998)). Additionally, by providing a tool for monitoring gene/transcript expression, the present invention is also useful for monitoring variations in gene/transcript expression in response to known mutations or polymorphisms in EMFs, or to identify previously unknown mutations or polymorphisms in EMFs based on variations in gene expression. Such polymorphisms in EMFs, particularly SNPs, may be useful diagnostic markers for disease.

As used herein, a sequence is said to "modulate the expression of an operably linked sequence" when the expression of the sequence is altered by the presence of the EMF. EMFs include, but are not limited to, promoters, and promoter modulating sequences (inducible elements). One class of EMFs is comprised of fragments that induce the expression of an operably linked gene/transcript in response to a specific regulatory factor or physiological event.

EMF sequences can be identified within the human genome by their proximity to the transcripts/exons provided by the present invention. An intergenic segment, or a fragment of the intergenic segment, from about 10 to 200, 10 to 500, 10 to 1kB, or 10 to 2.5kB nucleotides in length, preferably taken 5' from any one of the transcripts identified in the Sequence Listing (file SEQLIST.TXT), provided on the accompanying CD labeled CL001101CDA, will modulate the expression of an operably linked 3' gene/transcript in a fashion similar to that found with the naturally linked gene/transcript sequence. As used herein, an "intergenic segment" refers to fragments of the human genome that are between two transcripts herein described. Alternatively, EMFs can be identified using known EMFs as a target sequence or target motif in the computer-based systems of the present invention.

The presence and activity of an EMF can be confirmed using an EMF trap vector. An EMF trap vector contains a cloning site 5' to a marker sequence. A marker sequence encodes an identifiable phenotype, such as antibiotic resistance or a complementing nutrition auxotrophic factor, which can be identified or assayed when the EMF trap vector is placed within an appropriate host under appropriate conditions. As described above, an EMF will modulate the expression of an operably linked marker sequence. A sequence which is suspected as being an EMF is cloned in all three reading frames in one or more restriction sites upstream from the marker sequence in the EMF trap vector. The vector is then transformed into an appropriate host using known procedures and the phenotype of the transformed host in examined under appropriate conditions. As described above, an EMF will modulate the expression of an operably linked marker sequence.

Computer Related Embodiments

The nucleotide sequences provided by the present invention, a representative fragment thereof, or nucleotide sequences at least 99% identical to these sequences, may be "provided" in a variety of mediums to facilitate use thereof. As used herein, "provided" refers to a manufacture, other than an isolated nucleic acid molecule, that contains a nucleotide sequence of the present invention, i.e., the nucleotide sequences provided in the present invention, a representative fragment thereof, or nucleotide sequences at least 99% identical to these

sequences. Such a manufacture provides the coding portion of the human genome or a subset thereof (e.g., a human exon or transcript sequence) in a form that allows a skilled artisan to examine the manufacture using means not directly applicable to examining the human genome or a subset thereof as it exists in nature or in purified form.

In one application of this embodiment, a nucleotide sequence of the present invention can be recorded on computer readable media. As used herein, "computer readable media" refers to any medium that can be read and accessed directly by a computer. Such media include, but are not limited to: magnetic storage media, such as floppy discs, hard disc storage medium, and magnetic tape; optical storage media such as CD-ROM; electrical storage media such as RAM and ROM; and hybrids of these categories such as magnetic/optical storage media. A skilled artisan can readily appreciate how any of the presently known computer readable mediums can be used to create a manufacture comprising computer readable medium having recorded thereon a nucleotide sequence of the present invention. One such medium is provided with the present application, namely, the present application contains computer readable medium (CD-R) that has the transcript sequences provided/recorded thereon in ASCII text format in a Sequence Listing (provided in file SEQLIST.TXT on the accompanying CD labeled CL001101CDA.

As used herein, "recorded" refers to a process for storing information on computer readable medium. A skilled artisan can readily adopt any of the presently known methods for recording information on computer readable medium to generate manufactures comprising the nucleotide sequence information of the present invention.

A variety of data storage structures are available to a skilled artisan for creating a computer readable medium having recorded thereon a nucleotide sequence of the present invention. The choice of the data storage structure will generally be based on the means chosen to access the stored information. In addition, a variety of data processor programs and formats can be used to store the nucleotide sequence information of the present invention on computer readable medium. The sequence information can be represented in a word processing text file, formatted in commercially-available software such as WordPerfect and Microsoft Word, or represented in the form of an ASCII file, stored in a database application, such as OB2, Sybase, Oracle, or the like. A skilled artisan can readily adapt any number of data processor structuring formats (e.g., text file or database) in order to obtain computer readable medium having recorded thereon the nucleotide sequence information of the present invention.

By providing the nucleotide sequences of the present invention, a representative fragment thereof, or nucleotide sequences at least 99% identical to these sequences, in computer readable form, a skilled artisan can routinely access the sequence information for a variety of purposes.

Computer software is publicly available which allows a skilled artisan to access sequence information provided in a computer readable medium. Software which implements the BLAST (Altschul *et at, J. Mol. Biol.* 215:403-410 (1990)) and BLAZE (Brutlag *et at, Comp. Chem.* 17:203-207 (1993)) search algorithms on a Sybase system may be used to identify exons/transcripts within the human genome which contain homology to nucleic acid or proteins sequences from other organisms. Such exons/transcripts are protein-encoding fragments within the human genome and are useful in producing commercially important proteins such as therapeutic proteins.

The present invention further provides systems, particularly computer-based systems, which contain the sequence information described herein. Such systems are designed to identify commercially important fragments of the human genome.

As used herein, a "computer-based system" refers to the hardware means, software means, and data storage means used to analyze the nucleotide sequence information of the present invention. The minimum hardware means of the computer-based systems of the present invention comprises a central processing unit (CPU), input means, output means, and data storage means. A skilled artisan can readily appreciate that any one of the currently available computer-based systems are suitable for use in the present invention. Such system can be changed into a system of the present invention by utilizing the sequence information provided on the CD-R, or a subset thereof, without any experimentation.

As stated above, the computer-based systems of the present invention comprise a data storage means having stored therein a nucleotide sequence of the present invention and the necessary hardware means and software means for supporting and implementing a search means. As used herein, "data storage means" refers to memory which can store nucleotide sequence information of the present invention, or a memory access means which can access manufactures having recorded thereon the nucleotide sequence information of the present invention.

As used herein, "search means" refers to one or more programs that are implemented on the computer-based system to compare a target sequence or target structural motif with the sequence information stored within the data storage means. Search means are used to identify fragments or regions of the human genome that match a particular target sequence or target motif. A variety of known algorithms are disclosed publicly and a variety of commercially available software for conducting search means are available and can be used in the computer-based systems of the present invention. Examples of such software include, but are not limited to, MacPattern (EMBL), BLASTN and BLASTX (NCBIA). A skilled artisan can readily

recognize that any one of the available algorithms or implementing software packages for conducting homology searches can be adapted for use in the present computer-based systems.

As used herein, a "target sequence" can be any DNA or amino acid sequence of six or more nucleotides or two or more amino acids. A skilled artisan can readily recognize that the longer a target sequence is, the less likely a target sequence will be present as a random occurrence in the database. The most preferred sequence length of a target sequence is from about 10 to 100 amino acids or from about 20 to 300 nucleotide residues. However, it is well recognized that searches for commercially important fragments of the human genome, such as sequence fragments involved in gene expression and protein processing, may be of shorter length.

As used herein, "a target structural motif," or "target motif," refers to any rationally selected sequence or combination of sequences in which the sequence(s) is chosen based on a three-dimensional configuration which is formed upon the folding of the target motif. There are a variety of target motifs known in the art. Protein target motifs include, but are not limited to, enzymatic active sites and signal sequences. Nucleic acid target motifs include, but are not limited to, promoter sequences, hairpin structures and inducible expression elements (protein binding sequences).

A variety of structural formats for the input and output means can be used to input and output the information in the computer-based systems of the present invention. A preferred format for an output means ranks fragments of the human genome possessing varying degrees of homology to the target sequence or target motif. Such presentation provides a skilled artisan with a ranking of sequences which contain various amounts of the target sequence or target motif and identifies the degree of homology contained in the identified fragment.

A variety of comparing means can be used to compare a target sequence or target motif with the data storage means to identify sequence fragments of the human genome. For example, software which implements the BLAST and BLAZE algorithms (Altschul *et al., J Mol. Biol.* 215:403-410 (1990)) can be used to identify sequence fragments of interest within the human genome. A skilled artisan can readily recognize that any one of the publicly available homology search programs can be used as the search means for the computer-based systems of the present invention.

One application of this embodiment is provided in the figure. The figure provides a block diagram of a computer system 102 that can be used to implement the present invention. The computer system 102 includes a processor 106 connected to a bus 104. Also connected to the bus 104 are a main memory 108 (preferably implemented as random access memory, RAM)

and a variety of secondary storage devices 110, such as a hard drive 112 and a removable medium storage device 114. The removable medium storage device 114 may represent, for example, a floppy disk drive, a CD-ROM drive, a magnetic tape drive, etc- A removable storage medium 116 (such as a floppy disk, a compact disk, a magnetic tape, etc.) containing control logic and/or data recorded therein may be inserted into the removable medium storage device 114. The computer system 102 includes appropriate software for reading the control logic and/or the data from the removable storage medium 116 once inserted in the removable medium storage device 114.

The nucleotide sequences of the present invention may be stored in a well-known manner in the main memory 108, any of the secondary storage devices 110, and/or a removable storage medium 116. Software for accessing and processing the nucleotide sequence (such as search tools, comparing tools, etc.) reside in main memory 108 during execution.

All publications and patents mentioned in the above specification are herein incorporated by reference. Various modifications and variations of the described methods and systems of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has been described in connection with specific preferred embodiments, it should be understood that the invention as claimed should not be unduly limited to such specific embodiments. Indeed, various modifications of the above-described modes for carrying out the invention which are obvious to those skilled in the field of molecular biology or related fields are intended to be within the scope of the following claims.

Claims

That which is claimed is:

1) An isolated nucleic acid detection reagent that is capable of detecting the presence of 100,000 or more human exons, wherein said exons are selected from the group consisting of those identified in Table 1.

2) The detection reagent of claim 1, wherein said reagent is a nucleic acid array.

3) The array of claim 2, wherein said array is comprised of short oligonucleotides from about 5 to about 100 nucleotides in length.

4) The array of claim 2, wherein said array is comprised of polynucleotides based on the transcript sequences (SEQ ID NOS:1-39010), wherein said polynucleotides are from about 100 to about 1000 nucleotides in length.

5) An isolated nucleic acid detection reagent that is capable of detecting the presence of 2000 or more human exons, wherein said exons are selected from the group consisting of those identified in Table 1.

6) The detection reagent of claim 5, wherein said reagent is a nucleic acid array.

7) The array of claim 6, wherein said array is comprised of short oligonucleotides from about 5 to about 100 nucleotides in length.

8) The array of claim 6, wherein said array is comprised of polynucleotides based on the transcript sequences (SEQ ID NOS:1-39010), wherein said polynucleotides are from about 100 to about 1000 nucleotides in length.

9) An isolated nucleic acid detection reagent that is capable of detecting the presence of 5000 or more human exons, wherein said exons are selected from the group consisting of those identified in Table 1.

10) The detection reagent of claim 9, wherein said reagent is a nucleic acid array.

11) The array of claim 10, wherein said array is comprised of short oligonucleotides from about 5 to about 100 nucleotides in length.

12) The array of claim 10, wherein said array is comprised of polynucleotides based on the transcript sequences (SEQ ID NOS:1-39010), wherein said polynucleotides are from about 100 to about 1000 nucleotides in length.

13) An isolated nucleic acid detection reagent that is capable of detecting the presence of 10,000 or more human exons, wherein said exons are selected from the group consisting of those identified in Table 1.

14) The detection reagent of claim 13, wherein said reagent is a nucleic acid array.

15) The array of claim [redacted], wherein said array is comprised of short oligonucleotides from about 5 to about 100 nucleotides in length.

16) The array of claim 14, wherein said array is comprised of polynucleotides based on the transcript sequences (SEQ ID NOS:1-39010), wherein said polynucleotides are from about 100 to about 1000 nucleotides in length.

17) The detection reagent of claim 1, wherein said reagent is comprised of at least one polynucleotide spanning at least one exon-exon junction identified in Table 1.

18) The detection reagent of claim 5, wherein said reagent is comprised of at least one polynucleotide spanning at least one exon-exon junction identified in Table 1.

19) The detection reagent of claim 9, wherein said reagent is comprised of at least one polynucleotide spanning at least one exon-exon junction identified in Table 1.

20) The detection reagent of claim 13, wherein said reagent is comprised of at least one polynucleotide spanning at least one exon-exon junction identified in Table 1.

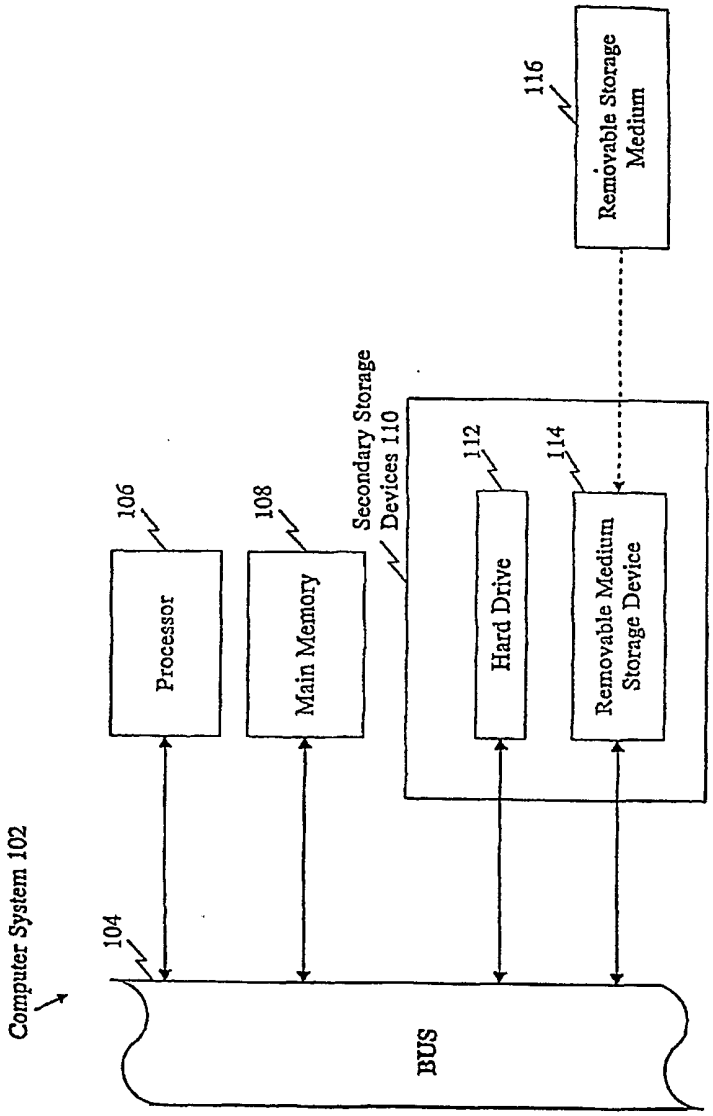

116
Removable Storage Medium
Secondary Storage Devices 110
112
114
106
108
Processor
Main Memory
Hard Drive
Removable Medium Storage Device
Computer System 102
104
BUS

FIGURE 1